USE OF SPECTRAL PEAK INFORMATION IN SPEECH RECOGNITION

M. Padmanabhan
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598
http://www.research.ibm.com/people/m/mukund

1 INTRODUCTION

In this paper, we address the issue of making use of spectral peak location information in a speech recognition system. The cepstral features that are used in most speech recognition systems, though perceptually motivated, do not explicitly model spectral peak trajectory information, which is a valuable clue to identifying the underlying phone. We present a study that examines the utility of using this information in speech recognition, to augment the information present in the cepstra.

We propose a method based on bandpass filtering the speech signal using several filters with different passbands, and using an adaptive IIR filter to track the locations of the spectral peaks in each bandpass output. This method has the advantage that along with the estimate of the spectral peak frequency, it also provides the energy at the spectral peaks (a feature that turns out to be quite informative). In initial experiments, the bandpass filters were chosen to correspond to the formant ranges, consequently, the locations of the spectral peaks are expected to correspond to the locations of the formants, for voiced sounds.

We next investigated the utility of using this spectral peak information to help discriminate between the phones used in speech recognition. In order to quantify the information provided by the new features (over and above the information provided by the cepstra), we measure the mutual information between the augmented feature vector (cepstra augmented with the new features) and the phonetic class labels, and compare it to the mutual information between the classes and the cepstra. Finally, we experimented with feature fusion techniques, where the new features were appended to the cepstra, and a new speech recognition system was trained on the augmented features.

2 ESTIMATION OF SPECTRAL PEAK FEATURES

We propose a method based on using an adaptive IIR filter for tracking the locations of the spectral peaks in the speech signal. In order to simplify the task of the adaptive filter, we first isolate spectral regions of the input signal by passing it through a bank of bandpass filters, such that each region contains only one dominant spectral component. In initial experiments, we based the choice of bandpass filters on physio-acoustical studies that indicate that the spectral peaks correspond to formant frequencies, and further, the first three formant frequencies lie in the range 280-710 Hz, 870-2250 Hz, 2250-2890 Hz [1]. Consequently, the speech signal is first filtered using a bank of three bandpass filters, with the passbands corresponding to these formant ranges, and subsequently, an adaptive IIR filter [2] is used to track the frequency at which the spectral energy is maximum within each passband. The frequency response of these bandpass filters is shown in Fig 1 - they are linear phase filters with all three filters having the same group delay.

As mentioned earlier, this method provides both the estimate of the spectral peaks as well as the energy at these peaks. For the voiced regions, this correlates roughly with the formant frequencies and energies, however, for the unvoiced regions, the adaptive filter essentially converges to the location of the spectral peaks in these regions. Further, as the adaptive filter is not allowed to change very abruptly, it also enforces a relatively smooth transition in the spectral peak locations over time.

2.1 Adaptive Filter

Denoting the outputs of the three bandpass filters as $y_1(t)$, $y_2(t)$ and $y_3(t)$, the adaptive filter stage identifies the spectral peak in the bandlimited spectra of

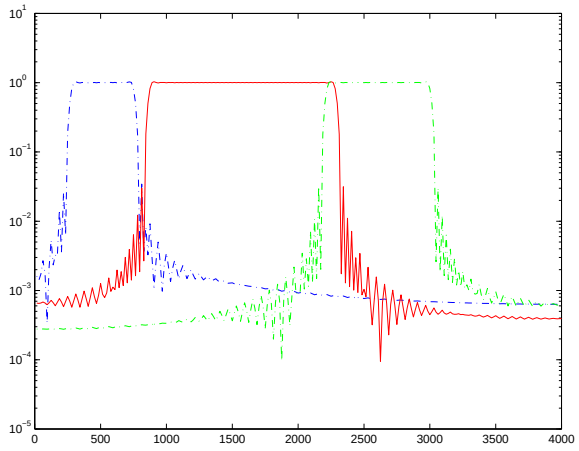

Figure 1: Frequency response of bandpass filters

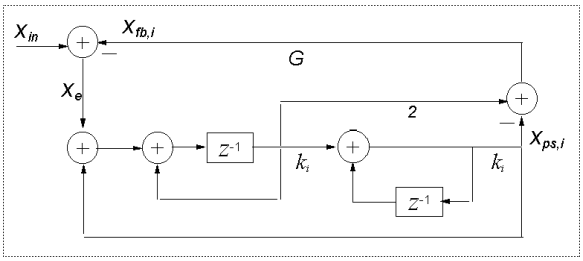

Figure 2: Adaptive filter structure

$y_i(t)$. This adaptive filter structure has previously been described in [2]. The basic idea is as follows: the adaptive filter is a multiple notch IIR filter with the notch frequencies being directly related to the filter coefficients (each notch frequency depends on exactly one filter coefficient). This multiple notch transfer function is obtained by embedding several digital resonators in a feedback loop, with the notch frequencies corresponding to the resonator frequencies. The goal of the adaptive algorithm is to minimize the power of the output of the notch filter. For the case where there are N sinusoids in the input, and there are N notches in the filter transfer function, the optimal solution is when the notch frequencies are equal to the input sinusoidal frequencies. In [2] an adaptive algorithm was described that guarantees converegence under certain conditions, and has complexity that is linear in N.

For our case, we assume that each bandpass output, $y_i(t)$ has a single spectral peak and use an adaptive filter with a single notch to track each of the $y_i(t)$. This filter structure is shown in Fig 2. The transfer function from the input to various nodes in the filter structure are given below:

$$H_e(z) = \frac{x_e}{x_{in}} = \frac{1-(2-k_i^2)z^{-1}+z^{-2}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (1)$$

$$H_{fb,i}(z) = \frac{x_{fb,i}}{x_{in}} = \frac{(2-k_i^2)z^{-1}-2z^{-2}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (2)$$

$$H_{ps,i}(z) = \frac{x_{s,i}}{x_{in}} = \frac{k_i^2 z^{-1}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (3)$$

The transfer function $H_e(z)$ represents a notch filter, with the notch frequency ω_i being related to the filter coefficient, k_i, through the following equation

$$k_i = 2\sin\left(\frac{\omega_i}{2}\right) \quad (4)$$

The algorithm for adapting the filter coefficient is given by

$$k_i(n+1) = k_i(n) - \mu \frac{x_e(n)x_{ps,i}(n)}{< x_{ps,i}(n)x_{ps,i}(n) > +\epsilon} \quad (5)$$

The term $x_e(n)x_{ps,i}(n)$ denotes the pseudo-gradient of the objective funtion (the coefficient is adapted in a direction opposite the pseudo-gradient), and the $< x_{ps,i}(n)x_{ps,i}(n) > +\epsilon$ denotes a power normalizing term that modfies the gradient direction to point in the Newton direction (i.e. the normalizing term approximates the inverse of the Hessian of the objective function). The spectral peak location can be inferred from the value of the filter coefficient k_i after it has converged using (4).

An additional feature of the filter structure is that the transfer function $H_{fb,i}$ is the complement of the notch transfer function, i.e., it represents a bandpass transfer function with the center frequency corresponding to the resonator frequency. Consequently, the power of the signal at $x_{fb,i}$ represents the power of the input speech signal at this frequency.

3 MUTUAL INFORMATION BETWEEN ACOUSTIC FEATURE AND PHONETIC CLASS

The usefulness of an acoustic feature may be measured by the amount of information it provides in discriminating between phonetic classes. This can be quantified by the mutual information between the feature vector and the phonetic class. Let c denote the phonetic class, and Z denote the acoustic feature vector. The mutual information between Z and c is defined by

$$I(Z;c) = \sum_c p(c) \int_Z p(Z/c) \log \left[\frac{p(Z/c)}{p(Z)}\right] dZ \quad (6)$$

Though (6) cannot be expressed in closed form, by vector quantizing Z and approximating the integral

with a summation, it may be rewritten as

$$I(Z;c) = \sum_{c} p(c) \sum_{Z_j} p(Z_j/c) \log \left[\frac{p(Z_j/c)}{p(Z_j)} \right] \qquad (7)$$

We are essentially interested in measuring the amount of information available by augmenting the usual cepstral feature vector with these new features. If Z represents the cepstral feature vector, and z represents one of the new features, then an augmented feature vector $\hat{Z}$ can be formed by concatenating Z and z. By vector quantizing $\hat{Z}$ into the same number of codewords as for Z, the mutual information between the augmented vector $\hat{Z}$ and c, $I(\hat{Z};c)$, can be computed from (8).

$$I(\hat{Z};c) = \sum_{c} p(c) \sum_{\hat{Z}_j} p(\hat{Z}_j/c) \log \left[\frac{p(\hat{Z}_j/c)}{p(\hat{Z}_j)} \right] \qquad (8)$$

The amount of information added by z can now be computed very simply

$$\delta I_z = I(\hat{Z};c) - I(Z;c) \qquad (9)$$

3.1 The amount of incremental information

The baseline acoustic feature, Z, was assumed to be the 13-dimensional Mel cepstral observation vector. We evaluated the δI_z associated with six new features, namely, the estimates of the first two formants from Xwaves [4], x_1, x_2, the spectral peak locations of the first and second adaptive filters, s_1, s_2 and the energy at these frequencies, e_1, e_2. The new features were observed either at the same time frame as the associated cepstra, or at any of the 4 preceeding or following time frames, leading to 54 sets of features. The classes, c corresponded to 58 phonetic classes. The number of feature vectors per class was limited to 20000, with the feature vectors for all classes totalling approximately 1 million. We K-means clustered the 1 million feature vectors to ≈ 2500 gaussians and used a likelihood metric to label each feature vector with the label of the closest gaussian. Subsequently, (8) was used to compute the mutual information between the class and the quantized feature vectors.

The unconditional class distribution had an entropy of 5.8 bits, and the mutual information associated with the cepstra alone, $I(Z;z)$ was 1.62 bits. Subsequently, we evaluated the incremental mutual information, δI_z, associated with each of the six new features at 9 time frames and plotted them in Fig 3.

Several observations may be made from Fig 3:
(i) the most informative features are the the energy at the peak frequencies, e_1, e_2. Further, the amount of additional information provided by this new feature is around 0.1 bits, which is not negligible when compared to the 1.62 bits of information that are provided by the cepstra.
(ii) the spectral peak estimates, s_1, s_2 provide less information than the energy
(iii) the formant estimates provided by Xwaves seem to add almost no information to the baseline cepstral feature. This is surprising because the formant estimates, x_1, x_2, should nominally provide the same amount of information as the spectral peak features, s_1, s_2. One possible reason for the discrepancy is that for the unvoiced segments of speech, the spectral peak locations tend to be influenced by the peak location in the preceding and following voiced segments and tend to vary smoothly between these values within the unvoiced segment, whereas the x_1, x_2 estimates vary more or less randomly for the unvoiced segments of speech. Consequently, even though the x_1, x_2 features do not provide information in disambiguating all phones, one could expect them to at least provide information to disambiguate speech regions with a well defined formant structure. To verify this conjecture, we computed the mutual information using (9), using data from only 8 voiced vowels, IY, IH, AE, AA, AH, UH and UW. The entropy of the prior distribution of these 8 classes is 3 bits, and the cepstra provide 1.03 bits of information. The incremental information provided by the new features is summarized in Fig 4. From this figure, it may be seen that the x_1, x_2 features do provide the maximum amount of information.

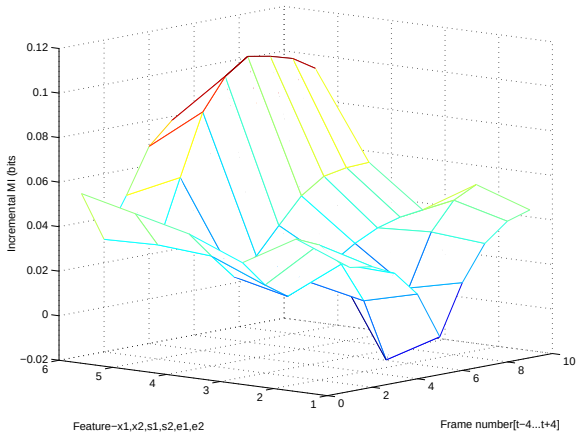

Figure 3: Incremental mutual information between new features and phonetic classes

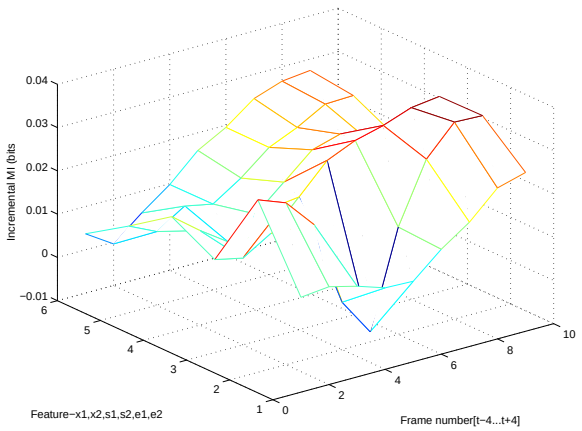

Figure 4: Incremental mutual information between new features and 8 vowels

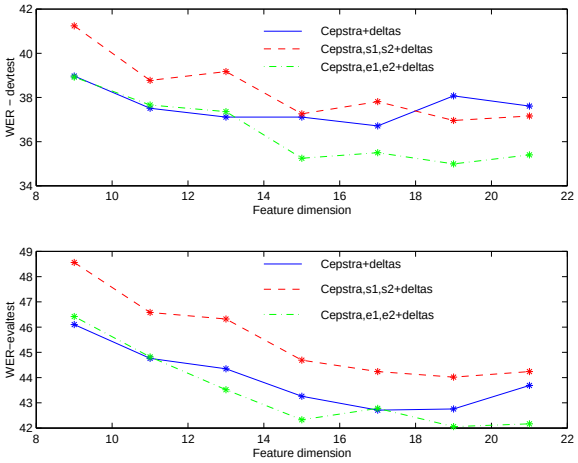

Figure 5: Word error rate vs feature dimension

4 SPEECH RECOGNITION

Next we experimented with feature fusion techniques to incorporate the new features into a speech recognition system. The task that we experimented on is a Voicemail transcription task [5], which represents large vocabulary spontaneous telephone speech. The size of the vocabulary is 15k words, and the perplexity of the trigram LM that was used is approximately 100. The amount of acoustic training data available is approximately 70 hours. We present results on both a development test set comprising of 43 voicemail messages (approximately 20 mts of speech) and an evaluation test set comprising of 86 voicemail messages (approximately 40 mts of speech).

4.1 System overview

The speech recognition system uses a phonetic representation of the words in the vocabulary. Each phone is modelled with a 3-state left-to-right HMM. Further, we identify the variants of each state that are acoustically dissimilar by asking questions about the phonetic context in which the state occurs. The questions are arranged heirarchically in the form of a decision tree, and its leaves correspond to the basic acoustic units that we model. An acoustic feature vector is extracted every 10 ms, and we model the pdf of the feature vector for each leaf of the decision tree, with a mixture of gaussians (our system had 2313 leaves and 34k gaussians). We experimented with using a varying number of cepstral features, and augmenting the cepstral feature space with the new features, s_1, s_2, e_1 and e_2. Subsequently, the final acoustic feature vector is obtained by augmenting the acoustic observation for a given frame with its first and second temporal derivatives.

4.2 Experimental Results

The word error rate results were computed on the Voicemail dev and eval test set [5] and are shown in Fig 5. The x-axis indicates the dimensionality of the extracted feature (either number of cepstra, or number of cepstra + e_1, e_2, or number of cepstra + s_1, s_2). The figure shows that the (e_1, e_2) estimates do contain more information than the higher order (13th) cepstra and can be used to improve the performance of the system (by 5.7% on the dev test and 5.2% on the eval test). The observation from Fig 4 that the (e_1, e_2) are more informative than the frequency estimates (s_1, s_2) themselves is also supported by the recognition results. However, there are some inconsistencies as well - for instance, Fig 4 indicates that adding s_1 and s_2 features to the cepstra should provide more information and help the recognition, however, this is not supported by Fig 5. Also, all the experiments conducted so far have been on relatively clean speech, possibly for noisy speech, the roles of energy and frequency estimate will be reversed, but this still remains to be established.

5 DISCUSSION

In this paper, we experimented with augmenting the cepstral features used in a speech recognition system with spectral peak related features. The new features were estimated by passing the speech signal through a bank of bandpass filters, and using an adaptive filter to track the location of the spectral peak in each band. We also quantified the amount of incremental information present in the new features by measuring the mutual information between the augmented feature and the phonetic classes and comparing it to

the mutual information between the cepstra and the phonetic classes. Finally we incorporated the additional features into a speech recognition system, using feature fusion, and showed that the new features contained more information that the higher order cepstra and could help improve the word error rate.

There are however several issues which are still open. It is not clear that the bandpass frequency ranges (motivated by formant frequency ranges) that were used for these experiments are the most appropriate. Possible experiments for the future include subdividing the entire frequency range into uniform intervals and tracking the spectral peak within each frequency range. Also, the simple feature fusion approach that was explored in this paper could probably be considerably improved upon by the use of alternative classifier combination schemes [6, 7].

6 ACKNOWLEDGEMENT

The author would like to thank Jun Huang for getting this work jump started during the summer of 1999, and Jing Huang and Karen Livescu for help with Xwaves.

7 APPENDIX

7.1 Comparison to Xwaves Formant Estimates

The bandpass filtering that preceded the spectral peak tracking was motivated by trying to isolate different formants, consequently, it would be interesting to examine the correlation between the spectral peaks and the formant frequencies for vowels which have a well defined formant structure. However, the true formant estimates are not available for any of the data that we experimented with, consequently, we used Xwaves [4] to estimate the locations of the formants from the speech signal. In Fig 6, we show the spectrogram of the speech for one sentence from a male and female speaker, as well as the first two formant estimates obtained from Xwaves (x_1, x_2), and the first two spectral peaks obtained from the adaptive filter (s_1, s_2). It may be seen that both the formant estimates provided by Xwaves and the spectral peaks tracked by the adaptive filter look reasonable, To validate the formant estimates that we obtained from Xwaves, we computed the average f_1 and f_2 frequencies for several vowels for one male and one female speaker, and plot them in Fig 7. Also plotted in the same figure are the average vowel positions for average American English, as specified in [1].

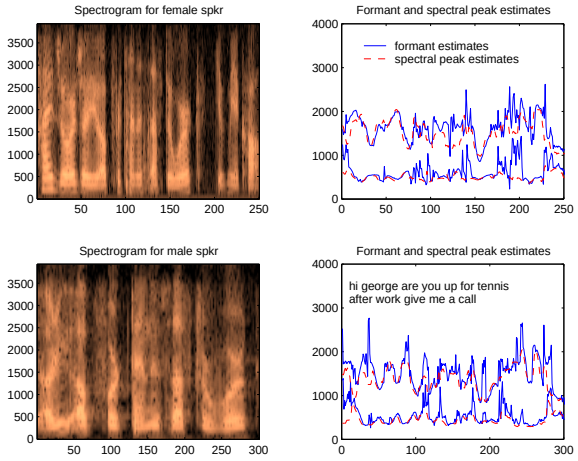

Figure 6: Spectrogram and formant frequency estimate of speech for one sentence with male and female speakers

We also computed the correlation between the Xwaves estimates and the adaptive filter estimates. The formant frequency estimates (x_1, x_2) were computed using Xwaves and the spectral peak estimates (s_1, s_2) using the adaptive filter, for 50 sentences from two speakers, one male and one female. Subsequently, we computed the mean and standard deviation of the differences $(x_i - s_i)$ and also the correlation between x_i and s_i for different phones. The correlation coefficient betwen x_i and s_i is defined as

$$\rho(x_i, s_i) = \frac{E[x_i s_i] - E[x_i]E[s_i]}{\sqrt{(E[x_i^2] - E[x_i]^2)(E[x_i^2] - E[x_i]^2)}}$$

The results are tabulated in Table 1. Here, $mean(x_i)$ indicates the mean value of the x_i for a given phone, $mean(x_i - s_i)$ and $var(x_i - s_i)$ indicate the mean and variance of the error between x_i and s_i for a given phone expressed as a percentage of the $mean(x_i)$, and $\rho(x_i, s_i)$ indicates the correlation between x_i and s_i. The statistics show a higher correlation between s_2 and x_2 than between s_1 and x_1. However, for unvoiced regions, the Xwaves estimates are more or less random, whereas the adaptive filter estimates are more consistent, in the sense that the spectral peak locations vary smoothly from the location in the preceding voiced segment. This enables them to be used in a simple feature fusion scheme, whereas initial experiments with using Xwaves estimates in feature fusion led to substantial degradation in recognition performance. [1].

[1] In the Xwaves package, we experimented with using various LPC orders to get the best agreement visually between a spectrogram and the formant estimates. The best compromise

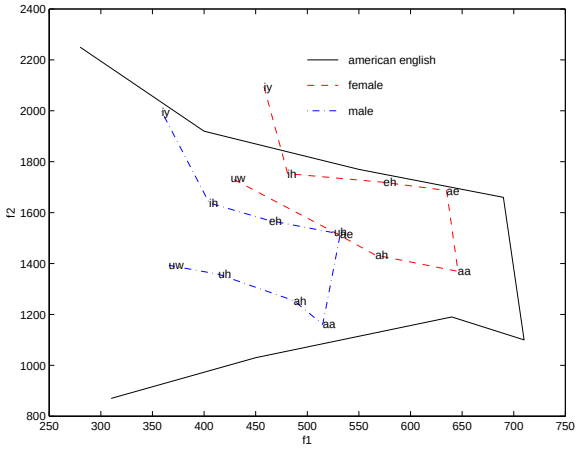

Figure 7: Location of vowels for male, female and average speakers

Phone	x_1	$mean(x_1 - s_1)$ % male/female	$var(x_1 - s_1)$ %	$\rho(x_1, s_1)$
IY	359/458	8/11	11/15	0.5/0.6
IH	405/481	13/11	18/13	0.2/0.3
EH	463/574	12/14	18/18	0.4/0.2
AE	532/635	16/22	18/16	0.5/0.4
AA	515/646	16/22	16/14	0.4/0.5
AH	487/566	14/15	20/28	0.4/0.1
UH	414/526	11/5	11/20	0.6/0.6
UW	366/426	9/5	30/19	0.5/0.3
	x_2	$mean(x_2 - s_2)$ % male/female	$var(x_2 - s_2)$ %	$\rho(x_2, s_2)$
IY	1997/2095	1/8	8/11	0.5/0.3
IH	1638/1753	5/14	10/10	0.8/0.7
EH	1569/1719	6/13	8/9	0.7/0.7
AE	1516/1687	10/18	12/10	0.6/0.6
AA	1161/1369	6/13	12/15	0.6/0.4
AH	1254/1433	7/15	7/13	0.9/0.6
UH	1358/1523	3/7	11/9	0.9/0.9
UW	1393/1736	0/4	19/7	0.8/0.9

Table 1: Comparison of statistics of Xwaves formant estimates (x_1, x_2) and spectral peaks (s_1, s_2)

7.2 Relationship between mutual information and heteroschedastic discriminant analysis

Given the reasoning in terms of mutual information that was used in this paper, it is interesting to note that heteroschedastic discriminant analysis (HDA) [8] can also be interpreted in terms of the mutual information between the projected feature vector and the classes. Let X represent the extracted acoustic feature, and Z represent a linear transformation of this feature, i.e., $Z = A\ X$. Our goal is now to find A such that the mutual information between Z and c is maximized. In (6) we defined the mutual information between the feature vector Z and class c. However, this expression may also be written as

$$I(Z;c) = H(Z) - H(Z/c) \tag{10}$$

Now, assume that the c^{th} class is modeled by a single full covariance gaussian in the X space, G_c, and that the mean and covariance of this gaussian is given by μ_c, Σ_c. Further, assume that the entire data is modeled with a single full covariance gaussian with mean and covariance μ, Σ [2]. Now the models for the complete data and the different classes in the Z space are also gaussians, with their means and covariances being given by $A\ \mu, A\ \Sigma\ A^T$ for the complete data, and $A\ \mu_c, A\ \Sigma_c\ A^T$ for the c^{th} class. Further, the self-entropy of a gaussian distribution can easily be derived to be $\frac{1}{2}log|\Sigma| + D + \frac{D}{2}log(2\pi)$. Consequently, (10) may be written as

$$I(Z;c) = log|A\ \Sigma\ A^T| - \sum_c P_c\ log|A\ \Sigma_c\ A^T| \tag{11}$$

which turns out to be almost exactly the objective function for the HDA as described in [8].

REFERENCES

[1] P. Ladefoged, "A Course in Phonetics", 1993, Harcourt Brace College Publishers, 301 Commerce Street, Suite 3700, Fort Worth, TX 76102.

[2] M. Padmanabhan and K. Martin, "Resonator-based filter-banks for frequency domain applica-

appeared to be an LPC order of 10, as hiher order LPC filters caused spurious peaks to appear in the power spectrum leading to a consistently lower estimate for f_1.

[2] Given that G_c represents the model of X conditioned on the class, c, one could also argue that the model for the complete data could be derived as $p(X) = \sum_c P_c p(X/c)$, where P_c represents the prior probability of class c. We will approximate this quantity by a single gaussian with mean μ and covariance Σ.

tions", IEEE Trans. Circuits and Systems, Oct 1991.

[3] S. M. Kay and S. L. Marple, "Spectral Analysis - A Modern Perspective", Proceedings of the IEEE, vol. 69, pp 1380-1419, 1981.

[4] Entropic Xwaves package.

[5] M. Padmanabhan et al., "Speech Recognition Performance on a VoiceMail Transcription Task", Proceedings of the Eurospeech 1999.

[6] L. Yuan and D. Miller, "Ensemble Classification by Critic-driven Combining", Proceedings of the ICASSP 1999.

[7] R. Schapire, Y. Freund, P. Bartlett, W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods", Annals of Statistics, 26(5): 1651-1686, 1998.

[8] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, "Maximum Likelihood Discriminant Feature Spaces", ICASSP 2000.